D-Cubicle: Boosting Data Transfer Dynamically for Large-Scale Analytical Queries in Single-GPU Systems

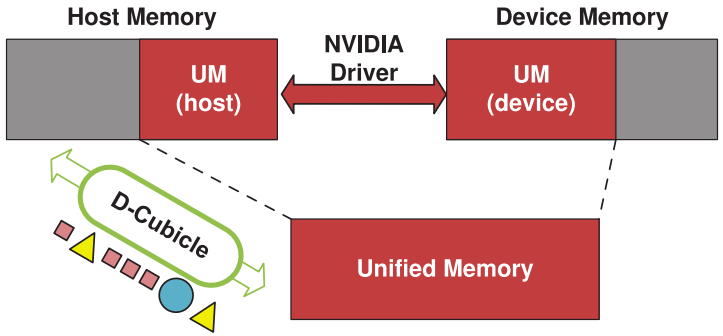
Jialun WANG, Wenhao PANG, Chuliang WENG, Aoying ZHOU

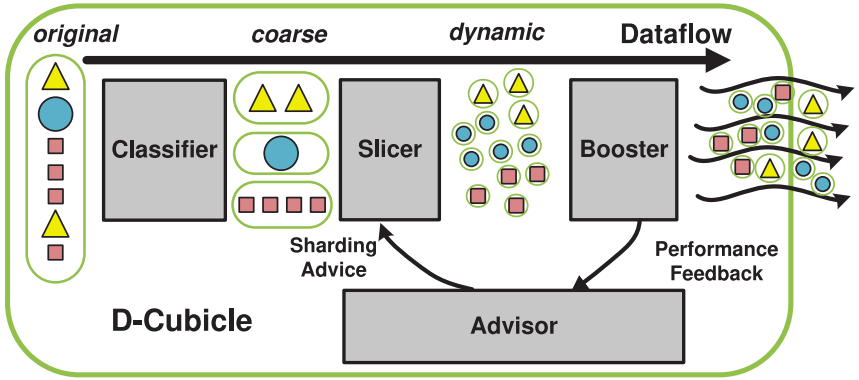
Frontiers of Computer Science, DOI: 10.1007/s11704-022-2160-z

Problems & Ideas

- Problems of transferring data between host and unified memory (UM) in GPU-accelerated large-scale analytical queries:
  - The actual host-UM transfer speed in analytical queries is much slower than the theoretical bandwidth, which slows down the overall performance heavily.

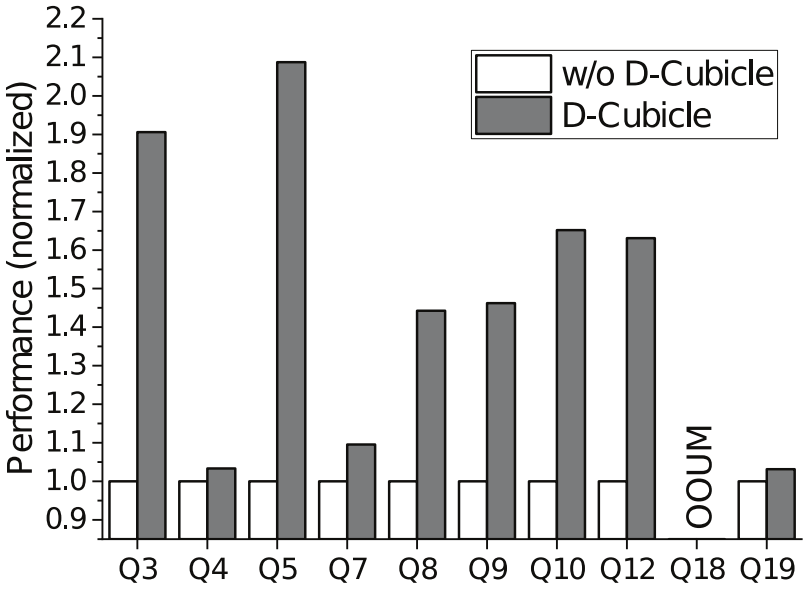
- Ideas: Slice the data and transfer them with multiple threads
  - Design a module to slice the data and accelerate the host-UM transfer with no need to change the query plans or introduce new operators.
  - Present a self-adaptive strategy to adjust the granularity of data and the number of threads according to the real-time performance feedback.



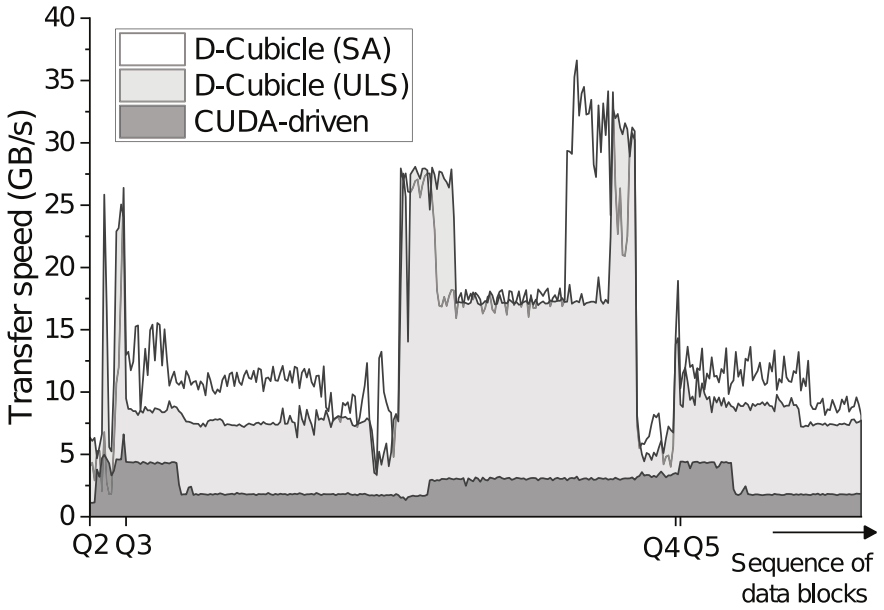


Main Contributions

- Performance of OmniSciDB with/without D-Cubicle (SF200)



- Real-time transfer speeds of self-adaptive mode (SA), static mode (ULS), and CUDA-driven mode



– D-Cubicle processes 200 GB of data on a single GPU with 32 GB of global memory, achieving **1.43x** averagely and **2.09x** maximally the performance of the baseline system.

– The self-adaptive method can adjust the strategy in real time and utilize the bandwidth better than the static method and the original CUDA-driven method.